26CS802 **Explainable AI** **4 -0 -0 -4**

Course Objectives:

This course provides an advanced understanding of the fundamental principles, need, and scope of explainability and interpretability in AI.

Course Outcome:

CO1	Describe the need and concepts of explainability in AI systems.
CO2	Apply intrinsic and model-agnostic techniques for interpreting machine learning models.
CO3	Implement explainability approaches for deep learning models.
CO4	Analyse AI systems for bias, fairness, and robustness.

Syllabus

UNIT I

Introduction to Machine Learning and Explainable AI, Purpose, need, and limits of explainability and interpretability, Explainability within the model development pipeline, Intrinsic vs. Post-hoc explainability Global and local explanations, properties and evaluation metrics

UNIT II

Intrinsic Explainable Models: Linear/Logistic Regression, Decision Trees, K-Nearest Neighbours. Model transparency and loss function behaviour. Model-Agnostic Techniques: Surrogate models, feature importance (global explanations) LIME, SHAP, Kernel Explainer, Integrated Gradients (local explanations) Bagging and Boosting explainability, SHAP for boosted trees.

UNIT III

Explainability in Deep Neural Networks: Post-hoc and agnostic approaches: adversarial features, data augmentation, occlusion methods. Layer-wise analysis: CAM, Grad-CAM, DeepSHAP, DeepLIFT. Counterfactual and causal inference concepts; What-If Tool, Contrastive Explanation Method (CEM). Temporal and Domain-Specific Explainability: Time-series and LSTM attribution using integrated gradients. Explainability for Transformers and Large Language Models (LLMs)

UNIT IV

Bias detection and mitigation in AI models. Fairness and ethical aspects of explainable systems. Adversarial robustness: evasion attacks, pre-processing defences, adversarial training. Evaluation and certification of robustness. Feature engineering and its role in interpretability. Practical applications and current research directions in XAI

Text Books

[1] Explainable AI with Python, Antonio Di Cecco and Leonida Gianfagna, Springer, 2021

[2] Christoph Molnar, Interpretable Machine Learning, Lulu Press, 2022.

[3] Ankur Taly et al., Explainable AI: Interpreting, Explaining and Visualizing Deep Learning, Springer, 2019.

Reference

[1] Hands-On Explainable AI (XAI) with Python: Interpret, visualize, explain, and integrate reliable AI for fair, secure, and trustworthy AI apps, Denis Rothman, Packt publisher, 2020

[2] Interpretable Machine Learning with Python: Learn to build interpretable high-performance models with hands-on real-world examples, by SergMasís , Packt publisher, 2021

Components	Continuous Assessment (Assignments/Presentation)	Internal Examination	Written Examination at the end of the semester
Weightage %	20	30	50